Reviewer Comments:

**Reviewer: 1**

Recommendation: RQ-Review Again after Major Changes

Comments:

This paper does a detailed investigation on overlapped speech detection and speaker counting, which is very important for the cocktail party problem. The analysis and survey on the human ability are also given, and this is nice for the people to do the basic comparison between the machine and human. I am not very familiar with this task, so I cannot evaluate the novelty of the proposed method. But I think the finding and analysis is very interesting and useful for the related community.

Other comments:

**1) It seems that there is no consistent corpus in this paper. From speaker counting to overlapped speech detection, and from machine to human. Different experimental setups are used. It is better to use the same data corpus to do the different evaluations in one paper, so people can follow the paper and do the comparison more easily.**

**FIXED**

**2) The proposed CNN-based deep learning method may be novel for this task, but actually is straightforward and commonly used for other speech tasks. So the author should clarify their contribution and novelty more clearly. Especially for the readers from other research areas.**

**3) Only the results from the author’s proposed method are given, but without any comparisons with other previous methods. Is this the first work and first deep learning method for this task? If not, it’s better to do the comparison with other state-of-art approaches.**

**4）The writing is easy to follow, but may not be clear enough for some parts. I think the author should improve the writing, and make it more concise in the revision.**

Additional Questions:

1. Is the topic appropriate for publication in these transactions?: Adequate Match

1. Is the paper technically sound?: Yes

2. How would you rate the technical novelty of the paper?: Somewhat Novel

3. Is the contribution significant?: Incremental

4. Is the coverage of the topic sufficiently comprehensive and balanced?: Treatment somewhat unbalanced, but not seriously so

5. Rate the Bibliography: Satisfactory

1. How would you rate the overall organization of the paper?: Could be improved

2. Are the title and abstract satisfactory?: Yes

3. Is the length of the paper appropriate? If not, recommend what should be added or eliminated.: Yes

null:

4. Are symbols, terms, and concepts adequately defined?: Yes

5. How do you rate the English usage?: **Needs improvement**

**Reviewer: 2**

Recommendation: RQ-Review Again after Major Changes

Comments:

I find it very interesting that this paper addresses the counting of simultaneous speakers from both a human performance and deep learning perspective, and I think both sets were well executed and well described in the paper. However, I found some serious issues in the paper that I think would need to be addressed prior to publication. Specifically, I am concerned with the way that results for different tasks on different datasets are being compared inappropriately, the fundamental task itself of counting speakers is not particularly well justified (especially for as many as 10 speakers), and there are sections of the paper that don’t seem to be relevant to the overall paper. I will address each of these three concerns in greater detail.

**For the first concern, there are several occasions (VII-B, VIII-A ) where the results of the paper’s deep network are compared against published numbers or the human results, but, as best as I can tell, these numbers were measured on completely different data. These comparisons are typically used to claim the presented method performs similarly well on shorter segments, but I don’t think such a claim can be made when it also could simply indicate that the network was tested on mixtures that happen to be easier. I think there needs to be a lot more care in making these comparisons. I also think there is a missed opportunity to measure the CNN performance on the actual human experiment data so that the comparison could be made in a more valid way. Even then, the fact that the CNN’s set of responses only goes up to 4 while the humans’ set of responses go up to 10 means that the tasks are fundamentally different, and that can’t be fixed by just ignoring the human trials with more than 4 speakers. Again, I think this needs to be treated with more care in the paper.**

**FIXED**

**Regarding my second concern, I think the problem of detecting overlapping speakers is well-justified in the introduction, but I don’t know why someone would want to count up to 10 simultaneous speakers (or even when that might reasonably occur). The CNN task is more justifiable, though, even there, it isn’t clear why the distinction between 2, 3, and 4 speakers is important. I think better justifying this task (instead of simply determining single speaker vs. multiple speakers) is critically important for the paper.**

**FIXED: Added in Section III**

**Third, there are several sections that don’t seem to fit in with the rest of the paper. First of all, Section III, regarding the previous strategy for counting, isn’t integrated at all with the other experiments (not to mention that the results given are overall error while all other results are broken down by number of speakers).** To some extent, this relates to my first concern, in that this issue could be resolved by measuring performance for this approach on the same dataset as the human experiments. This would be my recommendation, as I think showing performance on a common dataset for human performance, a previous method, and a new deep learning method would be the strongest integration of the various sections and result in the strongest paper. **But, short of that, I’m not sure how Section III fits in the paper.**

**FIXED: Removed section**

**Additionally, there is a discussion in the beginning of Section V about using features vs. raw samples along with another section in the bullets discussing raw samples as input, but I don’t think there are any results shown that utilize raw samples (except a histogram, which has its own description). I found that amount of discussion around an approach that didn’t appear to be used to be confusing and unnecessary.**

**FIXED – Removed the discussion**

I also had several smaller concerns:

**-Opening sentence: … presence < of > multiple … FIXED**

**-How was speaker familiarity ensured in the human experiments? FIXED - Explained**

**-Pg. 2, col. 2, line 45: there can be < a > maximum < of > REMOVED**

**-Pg. 5, col. 2, line 37: There are more variables changed between experiments than only language, so this claim seems to be too strong. FIXED**

**-Pg. 5, col. 2, line 48: If 4 is a reasonable upper bound, why did all the previous experiments go to 10? FIXED**

**-Pg. 6, col. 1, line 31: What is the relevance of word duration under stress here? This confused me. FIXED**

**-Pg. 7, col. 1, line 50: I wanted more details here about specifically when which combinations were added and when? Was this experimentally determined?**

**-In the experiments, in several places, the FFT is replaced with the spectrogram, but I did not see where the specifics of the FFT as features were defined. When it says FFT does that mean only one frame of the spectrogram? FIXED**

**-Fig. 13 should come before Fig. 12 FIXED**

**-Fig. 14 has both categorical accuracy and error. I would normally assume one to be one minus the other, but this doesn’t look to be exactly the case. What are these measurements?**

**-I think I would personally prefer Fig. 16 to be raw counts rather than percentage, so that comparisons across both rows and columns are meaningful, but that is just a suggestion. Not needed**

**-Pg. 10, col. 2, line 32: What supports the conclusion that single source recordings are more different than multi-source? This statement is very confident but I’m not sure how it was determined.**

**FIXED – Removed Comment**

Additional Questions:

1. Is the topic appropriate for publication in these transactions?: Adequate Match

1. Is the paper technically sound?: Yes

2. How would you rate the technical novelty of the paper?: Somewhat Novel

3. Is the contribution significant?: Moderately Significant

4. Is the coverage of the topic sufficiently comprehensive and balanced?: Yes

5. Rate the Bibliography: Satisfactory

1. How would you rate the overall organization of the paper?: Satisfactory

2. Are the title and abstract satisfactory?: Yes

3. Is the length of the paper appropriate? If not, recommend what should be added or eliminated.: Yes

null:

4. Are symbols, terms, and concepts adequately defined?: Yes

5. How do you rate the English usage?: Satisfactory

**Reviewer: 3**

Recommendation: RQ-Review Again after Major Changes

Comments:

The paper presents a comparison of humans and machines in the task of counting the number of overlapped speakers in a given speech segment. The presentation of the proposed work is clear enough.

The main concern is with the setup of using librispeech (read speech) spoken with close talking microphone for generating overlapped speech for human experiments. Even in single microphone recording conditions, **using a sum of close talking microphones is highly unnatural.** With the absence of spatial or temporal cues separating the speakers, the task also appears quite cumbersome. For example, for more than 5 speakers in the mix (Fig. 1 and Fig. 3), the listeners are operating at chance performance or less. The authors also mention “39% of the participants declared that they just guessed”. **The authors also find that the performance of the listeners does not improve over time. This brings the validity of the study into question. One may also argue whether humans indeed represent a good baseline for crowd sensing tasks based on audio as it is very artificial to have a large number of speakers speaking simultaneously with the similar energy for durations as long as 5-80s.**

**FIXED – Added new data to support the claim, and removed crowd-sensing as an application target.**

**The other major concern is the comparison between humans and machines. This comparison is flawed as humans were asked to make decisions between 1-10 talkers (10 class classification) while the machines are only making decisions between 1-4 speakers (4 class classification). The comment in Sec. VII.B “Compared to the performance achieved by human listeners, the CNN we trained achieves similar level of counting accuracy by analyzing a frame smaller by at least an order of magnitude” appears to be illformed. The authors mention in Sec. III.B for machine system building that “We picked 4 speakers as we considered it to be a reasonable upper limit for conversational speech”. This seems to be in direct contradiction with the human study and the motivation to use 10 speakers. The only possibility (if the human study is not to be repeated) for having a good human machine comparison would be to perform the same task in CNN system of classifying 1-10 speakers (10 class classification).**

**FIXED by new experiments**

**The discussion in Sec. VI needs some rewrite. For example, the statements like “there is a general belief that CNNs are faster by design …” comes across as unfounded and lacks scientific credibility.**

**FIXED**

**The machine performance is somewhat encouraging compared to similar other methods in [22],[57] where LSTMs were employed. However, the authors do not make an attempt to explicitly show that the proposed CNN models are better for this task compared to LSTMs. It is recommended that the authors implement the LSTM based systems ([57]) on the same dataset to highlight the improvements with the proposed system.**

**TODO? Claim that they get higher error in easier conditions with LSTM’s.**

**The motivation for the study in Sec. I is also lacking in some respect. The task of speaker counting discussed in the work is not directly related to selective auditory attention (SAA). ADDED Comments**

**It is not clear how the ability of counting competing speakers at a given time is important for tasks like speech recognition etc**. **TRUE, No SR. Added Comments.**

**Also, the authors give detailed description about overlapped speech detection which also has limited connection to speaker counting (the limited connection being one versus many task in overlapped detection, but most of the overlapped detection deals with single speaker versus two speaker overlaps).** A more precise introduction and motivation would serve well to appreciate the problem being solved.

**Partially fixed.**

Additional Questions:

1. Is the topic appropriate for publication in these transactions?: Adequate Match

1. Is the paper technically sound?: No

2. How would you rate the technical novelty of the paper?: Somewhat Novel

3. Is the contribution significant?: Incremental

4. Is the coverage of the topic sufficiently comprehensive and balanced?: Important Information is missing or superficially treated

5. Rate the Bibliography: Satisfactory

1. How would you rate the overall organization of the paper?: Could be improved

2. Are the title and abstract satisfactory?: Yes

3. Is the length of the paper appropriate? If not, recommend what should be added or eliminated.: Yes

null:

4. Are symbols, terms, and concepts adequately defined?: Yes

5. How do you rate the English usage?: Needs improvement